Estimation of a four-parameter item response theory model

By: Eric Loken, Kelly L. Rulison

This is the accepted version of the following article:

Loken, E. & Rulison, K. L. (2010). Estimation of a 4-parameter Item Response Theory model. *The British Journal of Mathematical and Statistical Psychology, 63(3),* 509-525. doi:10.1348/000711009X474502,

which has been published in final form at http://dx.doi.org/10.1348/000711009X474502.

Abstract:

We explore the justification and formulation of a four-parameter item response theory model (4PM) and employ a Bayesian approach to recover successfully parameter estimates for items and respondents. For data generated using a 4PM item response model, overall fit is improved when using the 4PM rather than the 3PM or the 2PM. Furthermore, although estimated trait scores under the various models correlate almost perfectly, inferences at the high and low ends of the trait continuum are compromised, with poorer coverage of the confidence intervals when the wrong model is used. We also show in an empirical example that the 4PM can yield new insights into the properties of a widely used delinquency scale. We discuss the implications for building appropriate measurement models in education and psychology to model more accurately the underlying response process.

Keywords: four-parameter item response theory model | delinquency scale | measurement models

Article:

1. Introduction

Item response theory (IRT) models are widely used in the social sciences. Although most early applications of IRT models were in education, applications now extend to other domains, including personality (Ferrando, 1994; Gray-Little, Williams, & Hancock, 1997; Reise & Waller, 1990; Rouse, Finger, & Butcher, 1999; Steinberg & Thissen, 1995), attachment (Fraley, Waller, & Brennan, 2000), psychopathology (Reise & Waller, 2003; Waller & Reise, 2009), attention deficit hyperactivity disorder (Lanza, Foster, Taylor, & Burns, 2005), and delinquency (Osgood, McMorris, & Potenza, 2002). IRT is clearly useful for scale construction and obtaining accurate latent trait estimates in multiple domains of inquiry.

Along with the growth in applications of IRT comes a need to consider carefully different parametric forms for IRT models, along with their interpretation and consequences for inferences. In this paper, we study a four-parameter model (4PM) that has received relatively little attention, as three more familiar models (see below), which specify the need for one, two, or three parameters to fit the data, dominate the literature. The 4PM we discuss allows each item's upper asymptote to be less than 1 (Barton & Lord, 1981; Hambleton & Swaminathan, 1985; Linacre, 2004; Rupp, 2003), to account for the possibility that even a very high ability respondent may on occasion answer an easy question incorrectly.

We believe that there are two reasons why the 4PM has not been widely discussed: first, the suggestions for its application have been rather isolated so that there is no clear consensus on the need for, or utility of, such a model (Barton & Lord, 1981; Hambleton & Swaminathan, 1985); and second, fitting the 4PM has been considered difficult, as even estimating the lower asymptote in the 3PM can be difficult (Baker & Kim, 2004; Embretson & Reise, 2000).

In what follows, we argue that there are several potential applications in education and psychology that may require a 4PM and we argue that it is fairly straightforward to fit the 4PM using a Bayesian approach. We then use a simulation study to demonstrate that inferences can be seriously compromised when the model used for analysis does not correspond to the data-generating model (i.e., when the 2PM or 3PM is used to analyse data generated from a 4PM). Finally, we demonstrate the application of the 4PM with an empirical example, analysing reports of delinquency in a large nationally representative study of youth.

2. IRT models

IRT models for binary response data usually assume a logistic curve for the probability of a 'correct answer' as a function of an underlying latent construct, ϑ. (We put 'correct answer' in quotation marks because one difficulty in moving back and forth between education and clinical applications of IRT is finding a vocabulary that bridges the domains. In psychopathology research, for example, the respondent is better described as 'endorsing an item' conditional on a certain level of a latent trait, as items generally do not have correct or incorrect answers. In what follows, we use both the educational and clinical vocabularies where appropriate without pausing again to emphasize their equivalence.) IRT models vary in how the functional relationship between ϑ and the response probability is represented. The one-parameter logistic model (1PM) assumes that all items have the same slope, and only differ in thresholds (item 'difficulty'). The two-parameter logistic model (2PM) allows items also to have different slopes and is consistent with evidence that not all items are equally discriminating (Ferrando, 1994; Gray-Little *et al.*, 1997; Reise & Waller, 1990). The three-parameter model (3PM, which is no longer technically in the logistic family) introduces a lower asymptote and is often used with multiple-choice items in educational testing, or for instruments where even low ability respondents have a finite probability of a correct response. Reference to a four-parameter model (4PM) has appeared sporadically in the literature (Barton & Lord, 1981; Hambleton & Swaminathan, 1985; Linacre,

2004; Rupp, 2003). However, this model has rarely been used in practice and until recently (see Osgood *et al.*, 2002; Reise & Waller, 2003; Tavares, de Andrade, & Pereira, 2004; Waller & Reise, 2009) was often dismissed as offering little practical benefit and as difficult to estimate. We address each of these issues in turn.

3. Justification and formulation of a four-parameter model

The first empirical investigation of a 4PM was by Barton and Lord (1981), who explored whether adding an upper asymptote less than 1 improved ability estimation on standardized tests. Their motivation was that the 3PM might be excessively punitive to high ability students who get an easy item incorrect (see also Mislevy & Bock, 1982). More specifically, the 3PM can accommodate a low ability student who correctly guesses a difficult item, but the upper asymptote of 1 in the 3PM assigns (effectively) a probability of zero that a high ability student incorrectly answers an easy item. To loosen that strong assumption, Barton and Lord estimated the following response model:

$$P(X_{ij} = 1|\theta_i; a_j, b_j, c_j, d) = c_j + (d - c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}}. \qquad (1)$$

Here, *aj* is the slope (or item discrimination) of the *j*th item, *bj* is the threshold, and *cj* is the lower asymptote. The upper asymptote, *d*, was fixed at 1 (yielding the 3PM), 0.99, or 0.98 and the *a*, *b*, and *c* parameters were fixed at values previously estimated using the 3PM. Barton and Lord re-estimated thousands of test scores for students taking three American academic exams, the Scholastic Aptitude Test, the Graduate Record Examination, and the Advanced Placement exam, and concluded that the changes in ability estimates with *d* set below 1 were too small to be of practical significance, especially given the difficulty of implementing the new model.

More recently, however, there has been a renewed interest in potential applications of a 4PM. For example, Osgood *et al.* (2002) analysed a self-report delinquency scale using IRT with multinomial responses. The 2PM graded response model provided good fit to the data, and yielded different information compared with the more straightforward total score on the scale. However, Osgood *et al.* also noted that there was always a chance that even the most delinquent youth would not report certain delinquent acts, and they suggested that future research should examine models that set an upper asymptote less than 1 (i.e., a 4PM model).

Within the field of psychopathology, Reise and Waller (2003) considered the need for an upper asymptote when modelling responses on the Minnesota Multiphasic Personality Inventory (MMPI). The MMPI is an assessment consisting of binary items where respondents endorse certain statements as true or not true of themselves. Although guessing is difficult to conceptualize for such an assessment, Reise and Waller found that some items appeared to require a lower asymptote greater than 0 because respondents very low on the underlying trait had a non-zero probability of endorsing the item. More importantly, when they reverse-coded their items (‘3PM-R’), they found that even more items required a non-zero lower asymptote,

suggesting that perhaps in the original keying these items should have been modelled with an upper asymptote less than 1. Reise and Waller (see also Waller & Reise, 2009) found that the overall model fit was not necessarily improved by moving to the 3PM, but that the test information function was dramatically different when the model was estimated as 2PM, 3PM, or 3PM-R. They concluded that new models with four parameters, representing both upper and lower asymptotes, should be considered for modelling some clinical and personality instruments.

Arguments in favour of a 4PM for IRT analyses have also appeared in genetics research. Tavares *et al.* (2004) modelled whether certain genes were activated or deactivated in individuals as a function of some quality of the person, such as a predisposition to an illness. The genes were considered as 'items' and, in general, individuals higher in the predisposition were more likely to have the genes activated. Because it is necessary to allow that low disposition individuals may still have the gene activated, and also that high disposition individuals may have the gene deactivated, Tavares *et al.* proposed a four-parameter model.

Although the need for a 4PM is becoming evident, there is some disagreement as to the exact form that it should take. The model given in(1), examined by Barton and Lord (1981), used a global *d* to represent a finite probability of carelessness across all items by all respondents. However, their modelling approach is not the most general implementation of the 4PM as they did not *estimate* the fourth parameter but rather fitted models with fixed values for *d.* In other applications, especially where the parameter is meant to capture a property of the item and not a tendency of the respondents, the upper asymptote may be item-specific. Reise and Waller (2003) give the example of an alienation scale item, 'Teachers dislike me'. The item is not universally endorsed by respondents high in alienation, suggesting the need for an upper asymptote less than 1. It is not likely, however, that all items from the alienation scale will share the same upper asymptote. Rouse *et al.*(1999) also argue that some psychopathology items might be viewed as undesirable, and even for respondents high in the assessed trait the probability of endorsement may be less than 1. Again, the degree of social desirability is likely to vary across items.

The more general formulation of the 4PM (Linacre, 2004; Rupp, 2003; Tavares *et al.*, 2004; Waller & Reise, 2009) suggests *d* as an item-specific upper asymptote implemented as:

$$P(X_{ij}=1|\theta_i;a_j,b_j,c_j,d_j)=c_j+(d_j-c_j)\frac{e^{1.7a_j(\theta_i-b_j)}}{1+e^{1.7a_j(\theta_i-b_j)}}. \qquad (2)$$

Even for respondents very high on the attribute, the expected probability of endorsing item *j* is *dj*<1, yielding an item-specific asymptote, as could be required for psychological assessment items, genetics applications, or other uses of the 4PM.

4. Estimating the 4PM with Bayesian methods

Another reason why 4PMs are rarely used is that these models have traditionally proved difficult to estimate using maximum-likelihood (ML) methods (Waller & Reise, 2009). Furthermore,

there has been a concern that estimates of *dj* would not be reliable, given the problems often encountered when estimating *cj* using ML (Baker & Kim, 2004; Embretson & Reise, 2000). We believe, however, that a Bayesian approach may be useful in achieving good estimation of the 4PM. Swaminathan and Gifford (1986) demonstrated that in comparison to ML methods, Bayesian estimation improved the reliability of the estimates of *cj* in the 3PM. More generally, Bayesian methods have proved to be very useful for estimating complex, heavily parameterized models where the likelihood is non-normal and multimodal, as may be the case when *d*is estimated as item-specific. Therefore, a Bayesian approach seems to be a convenient way to obtain reliable estimates of *dj*.

The Bayesian approach begins by defining the joint distribution of the parameters for *i* respondents, *j* items, and *ij* responses as

$$p(X, \theta, \phi) = \prod_i \prod_j p(X_{ij}|\theta_i, \phi_j) p(\theta_i) p(\phi_j). \qquad (3)$$

Here, φj represents the set of item parameters (*aj*, *bj*, *cj*, *dj*). The likelihood function $p(Xij|\vartheta i,\varphi j)$ is multiplied by the prior distribution of the item parameters, $p(\varphi j)$, and the prior distribution of the latent trait parameters, $p(\vartheta i)$, to give the full joint probability distribution. It follows by application of Bayes' rule that the posterior distribution $p(\vartheta i,\varphi j|X)$ is proportional to (3).

Estimation of the posterior means and variances is achieved through Markov chain Monte Carlo (MCMC) methods to simulate the full joint distribution (3), and thus also the marginal distribution of the parameters. The MCMC estimation can be conveniently carried out in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2004) where simulation of the posterior distribution is computed iteratively using a Gibbs sampler (Gelfand & Smith, 1990) to draw parameter estimates conditional on the previous draws of the other parameters. After a sufficiently long sequence of iterations, the draws should converge to the target distribution. An important consideration in MCMC simulations is determining whether the chain has converged to the target distribution. Evidence in favour of convergence includes stationarity of the time series with low autocorrelations in the marginal time series for each parameter, as well as good mixing of chains started from multiple initial values. We refer the reader to Gelman and Rubin (1992) for a more complete discussion of methods for assessing convergence.

The Bayesian approach requires specifying prior distributions for the parameters. Ideally, the choice of prior has a minimal impact on the posterior inference, and this is often true when there is a large amount of data. In the examples below, we use independent prior distributions commonly used in other applications in the literature.

A natural prior for the latent trait scores is the standard normal. In the absence of any additional information, respondent *i* is assumed to be randomly drawn from the population, usually considered normal with $M=0$ and $SD=1$. This is the standard prior used in software calculating the mode or mean of the posterior distribution of the trait given the data (Baker & Kim,

2004; Mislevy, 1986). A natural prior for the item difficulty parameters, *bj*, is also a normal distribution. However, although the item difficulties are supposed to be 'on the same scale' as the latent trait ϑ, the *bj* might be considered to come from a more diffuse normal distribution, such as *N*(0,2). A common prior for the slopes is to use a lognormal such as *p*(*aj*)=lognormal(0,0.125). A choice of prior for the *cj* is beta(5,17) (Mislevy, 1986). For the model under consideration, if we assume that the upper asymptote functions similarly to the lower asymptote, we can set *p*(*dj*)=beta(17,5). In the empirical example below, we also used a uniform prior on *d*, and found very similar results.

5. A simulation study

We first illustrate the Bayesian approach to estimating a 4PM through a simulation example that covers three different test lengths. As our purpose is to demonstrate the possibility of estimating the 4PM (along with a discussion of applications and properties), we do not provide an exhaustive evaluation of the model's properties under repeated sampling and multiple configurations. Instead, we examine three plausible scenarios that could be encountered in education or psychology. We then follow the simulated results with an analysis of empirical data.

Consider an instrument with *nj*=15, 30, or 45 items. A sample of *N*=600 respondents is generated from a normal population, *N*(0,1). The parameters of the items are drawn from the following parent distributions:

$$a_j \sim N(1.1, 0.3),$$

$$b_j \sim N(0, 1.1),$$

$$c_j \sim N(0.22, 0.05),$$

$$d_j \sim N(0.84, 0.05).$$

A 600 by *nj* response array was generated by having all the simulated respondents 'take' the *nj*-item instrument with probability of responding correctly given by (2). The response array of 0s and 1s ('incorrect' and 'correct') was then passed to WinBUGS to estimate the posterior means for the item and trait parameters (see Appendix for code). We ran WinBUGS for 25,000 iterations, discarding the first 10,000. The starting values for successive chains were randomly generated for *bj* and ϑ*i*, and were chosen to be 1.1 for *aj*, 0.2 for *cj*, and 0.8 for *dj*.

5.1. Parameter estimates of the 4PM

We first compare the posterior summaries for the item parameters with the values used to generate the data. Table 1 summarizes the correspondence between the true values and the estimates under the 4PM. In our sample of *N*=600 respondents, the correlation between posterior means of ϑ and the true scores was *r*=.80 for the 15-item test, and *r*=.92 for the 45-item test. The root mean square error (RMSE) for ϑ decreased from .58 in the shorter test to .39 in the longer

test. Across all three test lengths, the 95% credible intervals contained the true value 95% of the time. Because it is in the upper and lower tails of the distribution that test scores are often most relevant for inferences, we also tracked coverage for respondents with true ϑ greater than 1 or less than −1. For the 15-item test, the coverage in the tails was slightly below the advertised value (92%). Coverage was generally higher for the 30- and 45-item tests.

Table 1. Summary statistics for item parameter and trait $(\hat{\theta}_i)$ estimates: Four-parameter model

	Mean (SD)	Bias	RMSE	Proportion coverage of 95% CIs	Correlation with true value
a					
15 items	1.09 (0.11)	-0.005	0.26	1.00	.37
30 items	1.10 (0.12)	0.02	0.24	1.00	.48
45 items	1.11 (0.19)	0.009	0.20	.98	.66
b					
15 items	-0.06 (1.13)	-0.05	0.24	1.00	.98
30 items	-0.05 (1.15)	20.08	0.30	1.00	.97
45 items	-0.05 (1.09)	20.03	0.28	.98	.97
c					
15 items	0.22 (0.04)	0.001	0.05	1.00	.49
30 items	0.21 (0.04)	-0.006	0.05	.93	.54
45 items	0.21 (0.05)	-0.005	0.04	.98	.69
d					
15 items	0.81 (0.04)	0.04	0.06	.93	.64
30 items	0.81 (0.05)	-0.03	0.06	1.00	.52
45 items	0.81 (0.05)	-0.02	0.06	.87	.42
θ					
15 items	-0.001 (0.80)	0.04	0.58	.96 (.92/.92)[a]	.80
30 items	0.04 (0.89)	0.03	0.48	.95 (.95/.95)[a]	.88
45 items	0.004 (0.96)	0.04	0.39	.95 (.93/.98)[a]	.92

[a] Numbers in parentheses are the coverage for individuals with θ < -1 and θ > 1, respectively.

Jointly estimated with the person parameters were the item parameters. The *b* parameters showed little bias (−0.08 to −0.03), high interval coverage (≥98%), and high correlations between the estimated and true values ($r \geq .97$) across the three test lengths. The *a*, *c*, and *d*parameters also showed little bias and high interval coverage (≥87%). Unlike for *bj*, however, the correlation between the estimated and true values was lower (from .37 to .69). This probably reflects the small range of the values for these parameters compared to the standard errors of the estimates, a problem that is well known when estimating *cj* in the 3PM (Baker & Kim, 2004).

5.2. Consequences of estimating 4PM data using the 3PM and the 2PM

It is worthwhile to explore how the results would look should the 2PM and 3PM be used to analyse these data. In the 3PM, the upper asymptote is 1, implying that for a respondent high

enough on the trait the probability is essentially 1 that they will endorse easy items. The consequences of using the 3PM when there is a significant chance that the item is not endorsed should be predictable: unable to fit a function that rises quickly to 1, the slope estimates should be reduced, and the difficulty parameters should shift higher. The 2PM assumes lower and upper asymptotes of 0 and 1 respectively, and so when the data do not fit that assumption at either end, the slopes should be even further reduced from the 3PM.

Table 2 shows the item parameter estimates after fitting the 3PM using WinBUGS (which only requires altering the script in the Appendix by fixing all the *dj* to 1). As expected, the slopes are considerably lower; the mean slope is approximately 0.8 across the three test lengths, compared to the true slopes which had a mean of 1.10. At the same time, the item difficulties shifted higher by approximately half a standard deviation. Compared to the 4PM, the *cj* in the 3PM had slightly higher RMSE and slightly lower correlation with the true scores.

Table 2. Summary statistics for item parameter and trait $(\hat{\theta}_i)$ estimates: Three-parameter model

	Mean (SD)	Bias	RMSE	Proportion coverage of 95% CIs	Correlation with true value
a					
15 items	0.78 (0.17)	-0.32	0.44	.73	.22
30 items	0.79 (0.23)	-0.30	0.44	.63	.22
45 items	0.82 (0.26)	-0.28	0.42	.64	.32
b					
15 items	0.58 (0.86)	0.59	0.78	.40	.89
30 items	0.43 (0.98)	0.40	0.50	.57	.97
45 items	0.44 (0.94)	0.46	0.60	.53	.94
c					
15 items	0.24 (0.06)	0.02	0.07	.93	.26
30 items	0.22 (0.05)	-0.005	0.05	.93	.54
45 items	0.22 (0.06)	-0.002	0.05	.98	.60
θ					
15 items	0.00 (0.78)	0.04	0.60	.95 (.94/.82)[a]	.79
30 items	-0.006 (0.83)	-0.005	0.48	.95 (.96/.85)[a]	.87
45 items	-0.02 (0.86)	0.02	0.40	.91 (.94/.70)[a]	.91

a Numbers in parentheses are the coverage for individuals with $\theta < -1$ and $\theta > 1$, respectively.

Table 3 shows the comparable results after fitting the 2PM. The average slopes are now closer to 0.5. The *b* parameters have shifted back closer to 0. This probably reflects the need to accommodate the data at both ends of the trait distribution (note that the 3PM was able to bring the lower asymptote up from 0 with the *c* parameters).

Table 3. Summary statistics for item parameter and trait $(\hat{\theta}_i)$ estimates: Two-parameter model

	Mean (SD)	Bias	RMSE	Proportion	Correlation

				coverage of 95% CIs	with true value
a					
15 items	1.09 (0.11)	-0.005	0.26	1.00	.37
30 items	1.10 (0.12)	0.02	0.24	1.00	.48
45 items	1.11 (0.19)	0.009	0.20	.98	.66
b					
15 items	-0.06 (1.13)	-0.05	0.24	1.00	.98
30 items	-0.05 (1.15)	-0.08	0.30	1.00	.97
45 items	-0.05 (1.09)	-0.03	0.28	.98	.97
c					
15 items	0.22 (0.04)	0.001	0.05	1.00	.49
30 items	0.21 (0.04)	-0.006	0.05	.93	.54
45 items	0.21 (0.05)	-0.005	0.04	.98	.69
d					
15 items	0.81 (0.04)	0.04	0.06	.93	.64
30 items	0.81 (0.05)	-0.03	0.06	1.00	.52
45 items	0.81 (0.05)	-0.02	0.06	.87	.42
θ					
15 items	-0.001 (0.80)	0.04	0.58	.96 (.92/.92)[a]	.80
30 items	0.04 (0.89)	0.03	0.48	.95 (.95/.95)[a]	.88
45 items	0.004 (0.96)	0.04	0.39	.95 (.93/.98)[a]	.92

[a] Numbers in parentheses are the coverage for individuals with $\theta < -1$ and $\theta > 1$, respectively

Although the item parameters are different under the different models, this is an expected consequence of the limitations of the less parameterized models. The trait estimates for the respondents from both the 3PM and the 2PM correlated very highly (r>.98) with those from the 4PM, and there is effectively no bias in the estimates under any of the models. The posterior standard errors from both the 3PM and the 2PM were smaller than with the 4PM, but this then led to poorer coverage of the 95% intervals for ϑ, particularly in the tails of the distribution. For the 3PM, the confidence intervals were too wide for $\vartheta<-1$ and coverage was high. By contrast, in the upper tail ($\vartheta>1$) the intervals are too narrow and the interval coverage is much lower than 95%. For the 2PM, coverage in both tails of the distribution is impaired. In sum, although the individual trait estimates are almost identical under the three IRT models, the inferences are heavily affected by the application of the wrong model.

5.3. Consequences for the information functions

We further explore the issue of inference by considering the implications for measurement if the parameter estimates from the models described above were used in future applications of the instrument. For example, suppose that the instruments were to be used with these ‘known’ item parameters and that estimation of the trait scores were calculated using standard ML methods.

Confidence intervals would typically be constructed by first computing item information using the Fisher information function

$$I_j(\theta) = \frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)(1 - P_j(\theta))}. \qquad (4)$$

Test information is the sum of the information for the individual items under the assumption of conditional independence. The information function for the 4PM is

$$I_j(\theta) = \frac{(1.7a_j)^2(d_j - c_j)^2}{\left(c_j + d_j e^{1.7a_j(\theta - b_j)}\right)\left(1 - c_j + (1 - d_j)e^{1.7a_j(\theta - b_j)}\right)\left(1 + e^{-1.7a_j(\theta - b_j)}\right)^2}. \qquad (5)$$

The function reduces to the standard information functions for the 3PM and 2PM when all *dj* are set to 1 and all *cj* are set to 0, respectively. The test information functions comparing the 45-item 4PM to the 3PM, and 4PM to the 2PM, are shown in Figures 1 and 2, respectively.

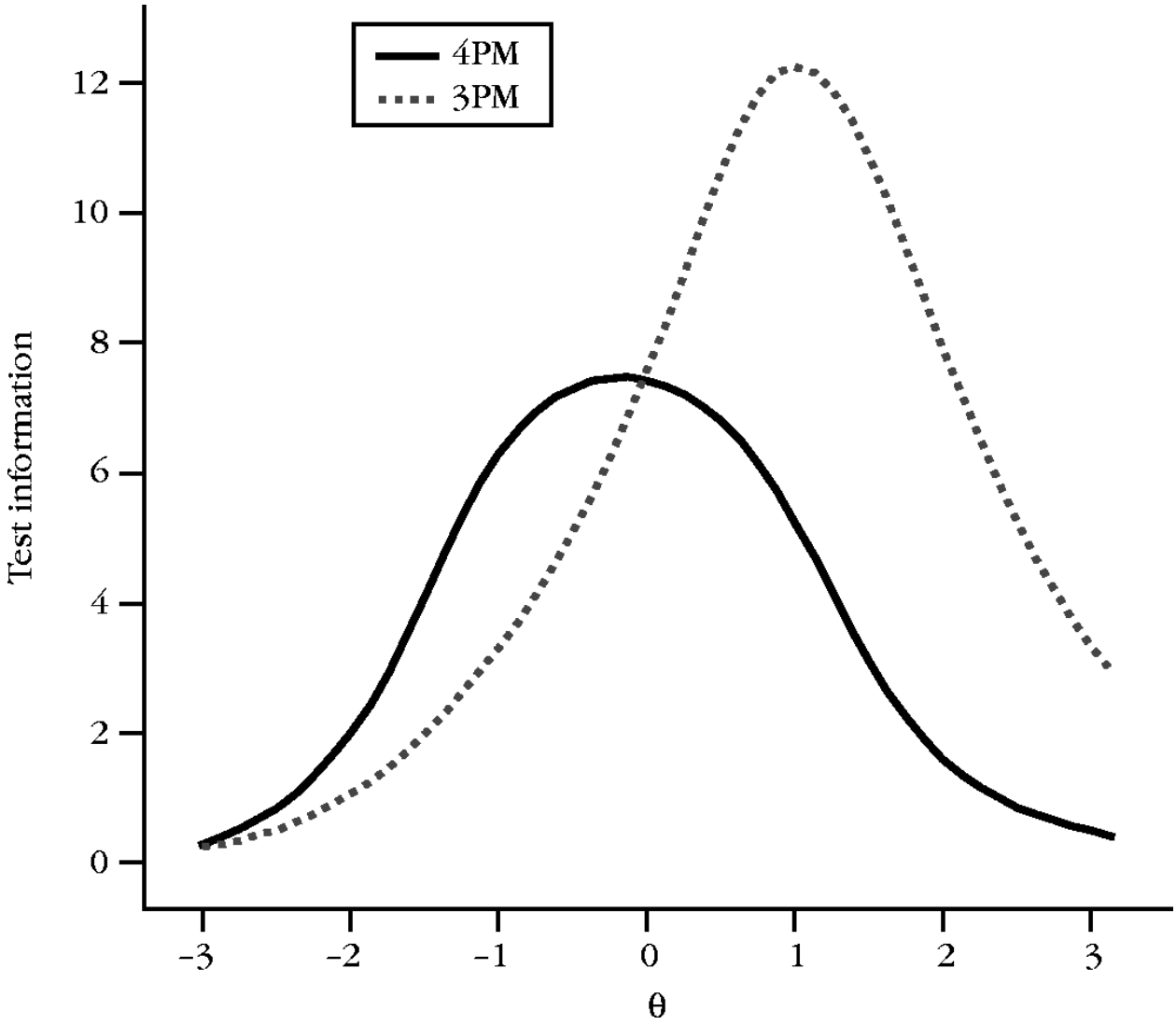

Figure 1. Test information functions for the 45-item simulated test for the 4PM and 3PM.

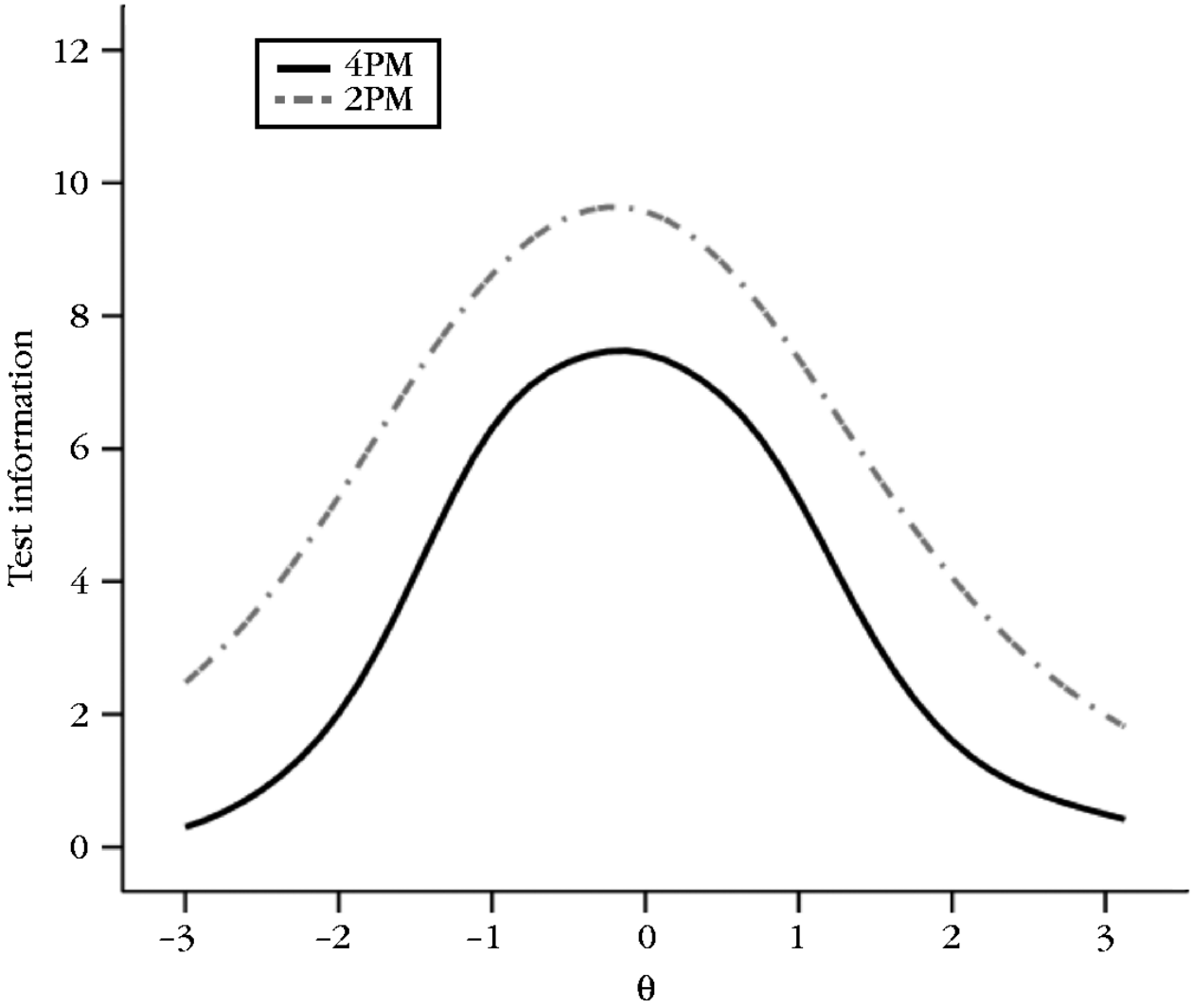

Figure 2. Test information functions for the 45-item simulated test for the 4PM and 2PM.

The effect of the parameter estimates in the 3PM (higher thresholds and lower slopes) is evident. At the lower end of the trait distribution, the function underestimates the information relative to the 4PM; at the upper end of the trait distribution, the information is severely overestimated. Inferences would be directly affected if the test were assumed to work according to the 3PM when the true response model was the 4PM. At the lower end of the trait distribution the standard errors would be too large (i.e., the 95% intervals would overperform); at the upper end of the trait distribution, the standard errors would be too narrow (i.e., the 95% interval would underperform).

The 2PM information function also shows the expected relationship relative to the 4PM. Although the *a* parameters are much lower for the 2PM, the overall information is still greater. The item difficulties are pulled to the middle, which stacks the information at the trait average.

Furthermore, because there is no discounting of information in the tails (no guessing at the low end, and no failures to endorse at the high end) the information in the tails is also assumed to be greater. The net result is that the assumed information function for the 2PM is higher than that for the 4PM.

5.4. Assessing model fit

We can also compare the overall fit of the two models. A feature of the MCMC estimation of the posterior distribution is that it is possible to calculate the fit at each iteration of the Gibbs sampler. Spiegelhalter *et al.* (2004) define the deviance information criterion (DIC) as a measure of fit analogous to information criteria such as the Akaike information criterion and Bayesian information criterion (Akaike, 1974; Schwarz, 1978). The posterior average of the deviance $-2 \log L$, where L is the likelihood function calculated at each iteration of the parameter draws, is augmented by the effective number of parameters. The effective number of parameters is estimated as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean for the parameters.

In the present example, the DIC clearly favoured the 4PM for the 30- and 45-items tests. For the 30-item test, the 4PM, 3PM, and 2PM DICs were 21,316, 21,336, and 21,394, respectively. For the 45-item test, the DICs were 31,918, 31,979, and 32,072, respectively. In both cases the DIC provides evidence in favour of the 4PM. For the 15-item test, the DIC favoured the 3PM over the 4PM (11,150 vs. 11,178). On the shorter test, for this one example, the increased model complexity was not justified by a sufficient improvement in fit.

6. Empirical example

We next demonstrate the utility of the 4PM with an empirical example. The data are responses to a self-report measure of delinquency from the 2005 Monitoring the Future (MTF) survey (Form 2) of 12th grade students (Johnston, Bachman, O'Malley, & Schulenberg, 2006). MTF is a national survey administered each year to about 50,000 8th, 10th, and 12th grade students to track trends in the attitudes and behaviours of American adolescents. As mentioned in the Introduction, an analysis of these data by Osgood *et al.* (2002) indicated that a model with an upper asymptote less than 1 might provide a better fit to the data, because even the most delinquent youth may not have committed some of the less serious offences in the last year.

On the survey, students reported the frequency with which they had engaged in 14 acts of delinquency within the past year (1=not at all to 5=five or more times). Although Osgood *et al.* (2002) retained all the response categories using the graded response model, they also questioned the value of the additional information provided by the more detailed frequency report as opposed to a simpler binary coding. We recoded the responses as 1 if the student reported engaging in the act at least once in the past year and 0 if they never engaged in that act. We limited our sample to students who provided complete data on the delinquency measure (N=2463; 96% of students who were administered this form).

The data were analysed with the same BUGS code as the simulations (the same start values were used). We found a similar pattern of results, although in this empirical example we did not have the benefit of knowing the 'true' model. Table 4 presents parameter estimates using the 4PM, 3PM, and 2PM. (The estimates can also be compared with Table 1 in Osgood *et al.*, 2002.) Allowing for an upper asymptote less than 1 led to much higher *a* parameters in the 4PM than in the 3PM and 2PM, as the line does not have to flatten out to accommodate the poorly fitting responses. The difficulty parameters (*bj*) are systematically lower in the 4PM for the same reason.

Table 4. Summary statistics for item parameter estimates by item: MTF

	4PM				3PM			2PM	
	a	b	c	d	a	b	c	a	b
1. Hit instructor	2.36	2.13	0.01	0.84	2.02	2.20	0.01	1.62	2.27
2. Serious fight	1.83	1.58	0.05	0.85	1.25	1.74	0.03	0.93	1.74
3. Gang fight	1.83	1.33	0.09	0.89	1.19	1.40	0.05	0.89	1.32
4. Hurt someone badly	2.08	1.47	0.04	0.86	1.35	1.62	0.02	1.09	1.61
5. Threaten with a weapon	2.55	2.08	0.01	0.87	2.27	2.11	0.01	1.65	2.20
6. Steal less than $50	3.96	0.39	0.03	0.74	1.21	0.85	0.02	1.10	0.82
7. Steal more than $50	1.64	1.53	0.01	0.82	1.72	1.62	0.01	1.63	1.60
8. Shoplift	3.36	0.47	0.02	0.72	1.27	0.92	0.02	1.16	0.89
9. Car theft	1.70	2.01	0.01	0.86	1.74	2.01	0.01	1.31	2.09
10. Steal car part	1.92	2.06	0.01	0.86	1.99	2.06	0.01	1.44	2.15
11. Trespass	1.33	0.86	0.07	0.83	1.06	1.07	0.05	0.88	0.98
12. Arson	2.11	2.12	0.01	0.85	2.09	2.13	0.01	1.72	2.18
13. School vandalism	1.56	1.29	0.02	0.76	1.26	1.56	0.01	1.15	1.54
14. Work vandalism	1.57	1.82	0.01	0.80	1.44	1.97	0.01	1.23	2.00

In this example, the *c* parameters are estimated to be very close to 0, suggesting that there may be very little 'false reporting' (analogous to guessing on a test). Nevertheless, c_3=0.09 for the question about being in a gang fight and c_{11}=0.07 for the item about trespassing. It is possible that some respondents who are low on the delinquency trait may answer yes to these questions, perhaps due to individual differences in the interpretation of these items. For example, some low delinquent youths may consider a scuffle between two groups of friends as a gang fight and there might be some latitude for interpretation of 'trespassing'.

The *d* parameters reflect a 15–20% chance that even highly delinquent youth will not report having committed specific delinquent acts within the last year. Delinquency items are not the same as academic test items that students solve in the act of taking a test. Rather, they are self-reports of behaviour over a fixed time period; even some highly delinquent youths will not have engaged in every act. Furthermore, to the degree that higher scores on the delinquency scale

reflect a developmental progression, some highly delinquent youths may have moved 'beyond' the mildest offences, and thus not committed these acts of delinquency within the past year.

The lowest *d* parameters were for the items about stealing things worth less than $50 (item 6), and shoplifting (item 8). For item 6, d_6=0.74 indicates that about a quarter of respondents do not endorse the item, regardless of delinquency level. This accommodates the observed empirical distribution of the responses, which did not have probability equal to 1 at the highest levels (for example, of the 25 people who endorsed eight of the 14 items, six *did not* endorse item 6). In a highly skewed distribution, such as a delinquency scale, the people at the highest level of the trait exert a lot of leverage and the model must accommodate them. Consider the parameters from the 4PM for item 6 (a_6=3.96 and b_6=0.39). If these were the parameters for the 2PM, a respondent with ϑ=2.0 would have a less than 1 in 50,000 chance of answering 'no' to having stolen something worth less than $50. But of the seven people who said 'yes' to 11 out of 14 items, two did not endorse this item. The gap between a predicted probability of 1 in 50,000 and an observed probability of 2 in 7 causes too much deviance, so the 3PM and 2PM flatten out the entire curve (e.g., under the 2PM, a_6=1.10 and b_6=0.82) to accommodate the outliers.

The 4PM, freed from the restriction of having to avoid reaching the asymptote too early, can have a steeper *a* and a lower *b*. In the case of items 6 and 8, the slopes are very steep (a_6=3.96; a_8=3.36). However, this closely matches the empirical data, because when respondents only endorse a few of the delinquency items, these are the items they are likely to endorse.

The dramatic differences in some of the *a* parameters across models suggest that the test information functions will differ. Figure 3 shows that the stealing and shoplifting questions have clearly made an impact on the 4PM information profile as their high discriminating power at moderate levels of the trait produces a visible bump. By contrast, the 2PM indicates that the test has far less information at moderate levels of the trait. The 2PM information function is very similar to that presented by Osgood *et al.* (2002) who discussed the lack of information of the MTF delinquency scale at moderate trait levels under the 2PM graded response model. Under the 4PM, the measure may be seen to carry more information than previously thought.

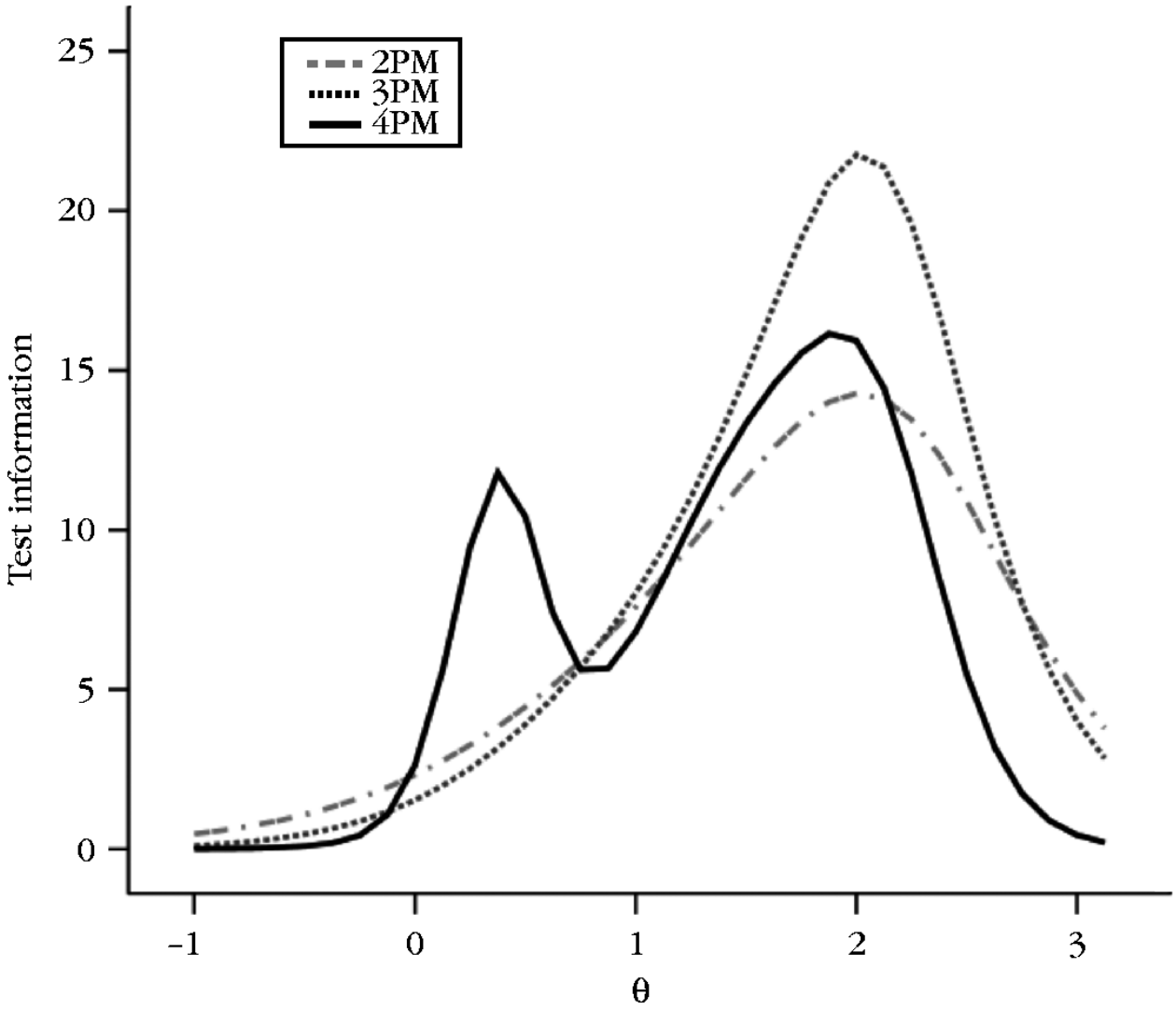

Figure 3. Test information functions for the juvenile delinquency scale for 4PM, 3PM, and 2PM.

Finally, in terms of model fit, we note that the DIC for the 4PM, 3PM, and 2PM was 16,865, 17,238, and 17,148, respectively, offering considerable support for the 4PM. We also analysed the data using a uniform prior on the *d* (uniform between 0.7 and 1.0) and found very similar results. Under a uniform prior, the DIC was even better (DIC=16,845), the *a*, *b*, and *c* parameters were almost the same, and the ϑ estimates and standard errors were essentially identical. The only difference was that under a uniform prior, the *dj* were higher. The choice of prior for *d* has some bearing on the parameter estimates, but in this case no bearing on the overall model choice, or on the essential insight of a different information profile when compared with more traditional IRT models.

7. Discussion

We have shown that it is possible to estimate an IRT model with item-specific upper asymptotes lower than 1. In our simulation, despite what might be considered a small sample of students

(*N*=600) for a complicated model, the MCMC algorithm converged and posterior estimates for both items and people closely reflected the parameter values used to generate the data. In our empirical example, we showed that the 4PM can provide improved model fit over standard approaches, along with new insights into the instrument's properties.

7.1. The need for a 4PM

We compared inferences between the 4PM, 3PM, and 2PM for data generated under the 4PM. Although the ϑ estimates were highly congruent (r>.98) across all approaches, there were differences in the quality of inference. Using the 3PM, standard errors at the lower end of the trait distribution were much larger than when the correct model was used, and the standard errors at the upper end of the trait distribution were much smaller. When the 2PM was used to analyse data generated under 4PM conditions, the inference was affected symmetrically, with overly narrow standard errors at both ends of the trait distribution.

The difference in the information functions among the models is likely to be of practical significance. Our analysis of the delinquency data from the MTF study showed that under the 4PM there might be more information about moderate levels of delinquency than previously believed. In particular, two items were seen to be highly discriminating when $0<\vartheta<1$ once the assumption of unanimous endorsement at higher levels was relaxed.

In clinical samples, the high and low trait populations are often of interest (Reise & Waller, 2003). As other researchers have noted (Stark, Chernyshenko, Drasgow, & Williams, 2006), personality assessment items are often written at the extremes of the scale. If items have an appreciable chance of not being endorsed, even by the respondents highest on the trait, then analysing the instrument with either 2PM or 3PM can result in distorted inference.

In one recent example, Rouse *et al.* (1999) applied the 3PM to data from the MMPI PSY-5 scales. The neuroticism scale (NEM) *b* parameter estimates averaged 1, and the *a* parameter estimates were very low. Rouse *et al.* present a 3PM test information function for the NEM (p. 301) which suggests that the test performs best at the high end of the distribution; however, our results in Figures 1 and 3 raise the possibility that the 3PM may overestimate test information at the high end. Even more striking, item estimates for the psychoticism subscale had extremely high *b* and very high *a*. If these items were generally unlikely to be endorsed, and not uniformly endorsed by respondents at the high end of the distribution, then such parameter estimates are a predictable result. (A further complicating factor in Rouse *et al.* is that the *b*parameters may have been capped at 3.0.) It would be interesting to see if a 4PM analysis might yield different insights.

Another recent clinical example indicates that an adolescent low self-esteem scale might require an upper and lower asymptote (Waller & Reise, 2009). These authors discussed a series of reasons for adding an upper asymptote, including asymmetric ambiguity and underlying multidimensionality. An example of asymmetric ambiguity is when an item may strike

respondents at one end of the trait spectrum as ambiguous, thus influencing their responses. For example, a student with low self-esteem who is asked whether friends have made him do bad things may disagree with the item not because he is confident, but because he rarely interacts with friends. If a sufficient proportion of respondents with very low self-esteem confronted the item in the same way, the model would need to accommodate the mixture of responses. In the same way that the *c* parameter reflects a mixture of guesses and 'real' answers, the *d* parameter can reflect a mixture of interpretations of an item at the upper end of the spectrum.

In educational testing, the need to model an upper asymptote below 1 generally seems less necessary. In high-stakes testing, the student is assumed to be maximizing her utility, and therefore if she can answer the question correctly she will do so. It was more than 25 years ago that Barton and Lord (1981) discarded the idea. However, we have shown elsewhere that lowering the upper asymptote reduces the bias in computerized adaptive testing algorithms when high ability students make mistakes early in the test (Rulison & Loken, 2009). We believe that there will be other educational testing contexts where the Barton and Lord idea might prove useful, including modelling responses in non-high-stakes testing situations (such as practice tests or computer adaptive study materials).

7.2. Alternative formulations of the 4PM

These educational testing examples also lead us to consider alternate forms of a 4PM. Barton and Lord's (1981) formulation was given in equation (1), where the purpose of *d* was to designate a propensity to respond incorrectly. However, this interpretation refers to a tendency in response behaviour, not to the quality of the item. The potential for a respondent to be careless might be better characterized by modelling the observed responses as a mixture of careless and non-careless responses. If we let *d* designate the proportion of non-careless responses, then the proper description of the function would be

$$P(X_{ij}|\theta_i; a_j, b_j, c_j, d) = (1-d)c_j + d\left(c_j + (1-c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1+e^{1.7a_j(\theta_i - b_j)}}\right)$$

$$= c_j + d(1-c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1+e^{1.7a_j(\theta_i - b_j)}} \qquad (6)$$

Equation (6) may be a more accurate representation of what Barton and Lord (1981) intended to implement. This representation further raises the question of whether *d* could be considered as a fixed characteristic of the test situation, or as an attribute of a person. The latter possibility could require that a fourth parameter be estimated as person-specific or it could suggest that another approach is required altogether, such as the use of a biweight estimator (Mislevy & Bock, 1982).

Yet another possible adjustment to the model is to consider Reise and Waller's (2003) suggestion that a test might be composed of a mixture of items, some requiring upper and some requiring lower asymptotes. Such a model could be achieved by specifying the response function as a mixture; it might also be achieved by loosening the prior distribution on the *c* and *d* parameters

to accommodate a wider range of values (i.e., some, but not all, *c* close to 0 and some *d* close to 1). We believe that these models are estimable but that some computational and identification issues would need to be addressed.

7.3. Summary

We recognize that much more work needs to be done on estimating and interpreting the 4PM. Our simulation was a demonstration of one plausible approach to estimating what is considered a challenging model. We were able to obtain good estimates with a modest sample size of 600 (a sample considered minimal even for good ML estimation of the 3PM), but the degree of consistency under different conditions and different choices of prior distribution for the *c* and *d* parameters requires systematic exploration. In our empirical example, we provided a new look at a popular delinquency scale. However, more empirical examples will be required in the future to further establish the utility of the 4PM.

Although we used the DIC here to report a global index of fit with penalty for model complexity, more can be done to investigate model fit, especially in more thorough simulations. The DIC can be used to check for the relative contributions to the misfit from the various parameters. We should also consider standard measures of item and person fit as diagnostic and model selection criteria.

Although we took a Bayesian approach in this study, we do not mean to imply that fitting a 4PM necessarily requires a Bayesian approach. As a reviewer points out, the Bayesian approach incorporates (hopefully minimally impactful) amounts of prior information that facilitates estimation. It remains an open question whether a ML approach with some (hopefully minimally impactful) constraints could achieve more efficient estimation of the models considered in this paper. We see this as one of many promising areas of future work.

In sum, we believe that there is a need in psychological measurement (both in education and other assessment fields) for item response models that relax the assumption that the probability of a correct response necessarily goes to 1 when the respondent's trait level is sufficiently greater than the difficulty of the item. We have shown one straightforward way to estimate such models using a Bayesian framework, and we have also shown the consequences of analysing data generated under a 4PM performance model with either a 2PM or 3PM.

Acknowledgements

Support for this research was provided by NSF award SES-0352191 (PI Loken) and the National Institute on Drug Abuse (DA 017629; DA 024497-01). The authors would like to thank Niels Waller and Wayne Osgood for very helpful discussions.

Appendix

WinBUGS code for 4PM

```
model
 {
 for (i in 1: nstud) {
 for (j in 1: nqs) {
p[i, j]<- c[j]+(d[j]−c[j])*(exp(1.7*a[j]*(theta[i]−b[j]))/(1+exp(1.7*a[j]*(theta[i]−b[j]))))
 r[i, j]~dbern(p[i, j])
    }
    theta[i]~dnorm(0,1)
   }
   for (k in 1:nqs) {
    a[j]~dlnorm(0,8);
    b[j]~dnorm(0,.25);
    c[j]~dbeta(5,17);
    d[j]~dbeta(17,5);
 }
 }
```

References

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automated Control* **19** 716–723. 1974.

Baker, F. B., Kim, S. *Item response theory: Parameter estimation techniques*. 2nd ed. New York Marcel Dekker 2004.

Barton, M. A., Lord, F. M. *An upper asymptote for the three-parameter logistic item-response model* Princeton, NJ Educational Testing Service 1981.

Embretson, S. E., Reise, S. P. *Item response theory for psychologists* Mahwah, NJ Erlbaum 2000.

Ferrando, P. J. Fitting item response models to the EPI-A impulsivity subscale. *Educational and Psychological Measurement* **54**118–127. 1994.

Fraley, R. C., Waller, N. G., Brennan, K. A. An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology* **78** 350–365. 2000.

Gelfand, A. E., Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85** 398–409. 1990.

Gelman, A., Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–511. 1992.

Gray-Little, B., Williams, V. S. L., Hancock, T. D. An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin* **23** 443–451. 1997.

Hambleton, R. K., Swaminathan, H. *Item response theory: Principles and applications* Norwell, MA Kluwer Academic Publishers 1985.

Johnston, L. D., Bachman, J. G., O'Malley, P. M.,&Schulenberg, J. E. (2006). *Monitoring the Future: A continuing study of American youth (12th-grade survey), 2005* [Computer file]. Inter-University Consortium for Political and Social Research conducted at the University of Michigan, Institute for Social Research, Survey Research Center, ICPSR04536-v1, Ann Arbor, MI.

Lanza, S. T., Foster, M., Taylor, T. K.,&Burns, L. (2005). *Assessing the impact of measurement specificity in a behavior problems checklist: An IRT analysis*. Technical Report 05-75. University Park, PA: The Pennsylvania State University, The Methodology Center. Retrieved from methodology.psu.edumediabibliographytechreports197934924805-75.pdf.

Linacre, J. M. Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions* **18**959–960. 2004.

Mislevy, R. J. Bayes model estimation in item response models. *Psychometrika* **51** 177–195. 1986.

Mislevy, R. J., Bock, R. D. Biweight estimates of latent ability. *Educational and Psychological Measurement* **42** 725–737. 1982.

Osgood, D. W., McMorris, B. J., Potenza, M. T. Analyzing multiple-item measures of crime and deviance I: Item response theory scaling.*Journal of Quantitative Criminology* **18** 267–296. 2002.

Reise, S. P., Waller, N. G. Fitting the two-parameter model to personality data. *Applied Psychological Measurement* **14** 45–58. 1990.

Reise, S. P., Waller, N. G. How many IRT parameters does it take to model psychopathology items?. *Psychological Methods* **8**. 2.164–184. 2003.

Rouse, S. V., Finger, M. S., Butcher, J. N. Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment* **72** 282–307. 1999.

Rulison, K. L., Loken, E. I've fallen and I can't get up: Can high ability students recover from early mistakes in computer adaptive testing?. *Applied Psychological Measurement* **33** 83–101. 2009.

Rupp, A. A. Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing* **3**. 4. 365–384. 2003.

Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **6** 461–464. 1978.

Spiegelhalter, D. J., Thomas, A., Best, N. G.,&Lunn, D. (2004). *WinBUGS user manual* (Version 2.0). Cambridge: MRC Biostatistics Unit.

Stark, S., Chernyshenko, O. S., Drasgow, F., Williams, B. A. Examining assumptions about item responding in personality research: Should ideal point methods be considered for scale development and scoring?. *Journal of Applied Psychology* **91** 25–39. 2006.

Steinberg, L., Thissen, D. Item response theory in personality research Shrout, P. E., FiskeS. T. *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* 161–181. Hillsdale, NJ Erlbaum 1995.

Swaminathan, H., Gifford, J. A. Bayesian estimation in the three-parameter logistic model. *Psychometrika* **51** 589–601. 1986.

Tavares, H. R., Andrade de, D. F., Pereira, C. A. Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology* **27** 679–685. 2004.

Waller, N. G., Reise, S. P. Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPIEmbretson, S., RobertsJ. S. *New directions in psychological measurement with model-based approaches* Washington, DC American Psychological Association 2009.